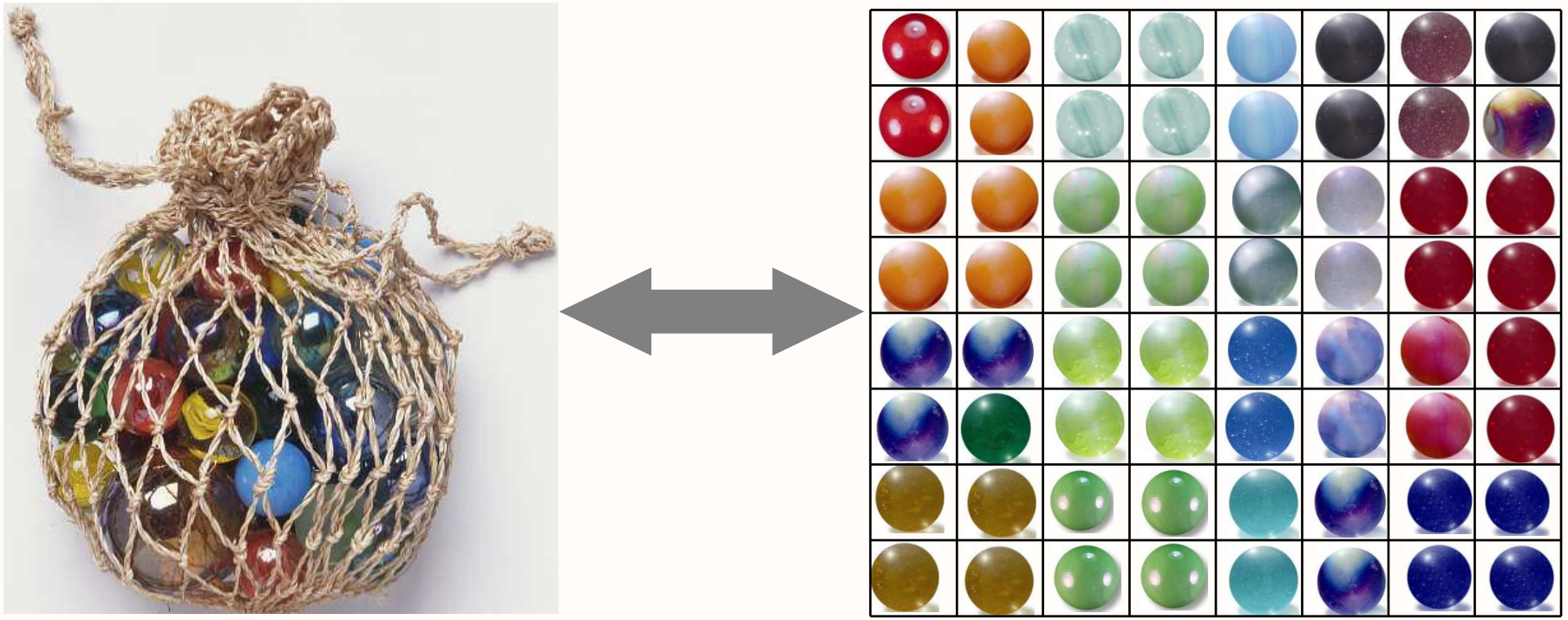
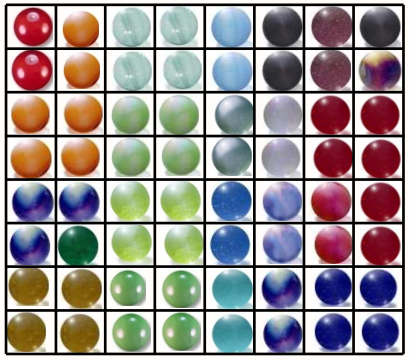
Geometry and Molecular Information



von Neumann – Lederberg symposium 2009

Geometry of molecular information channels



"Marble packing"

(A) Max colors.

(B) Same\similar color of neighbors.

Channel

Mapping

Rate

Distortion

Fitness

Transition

Topology

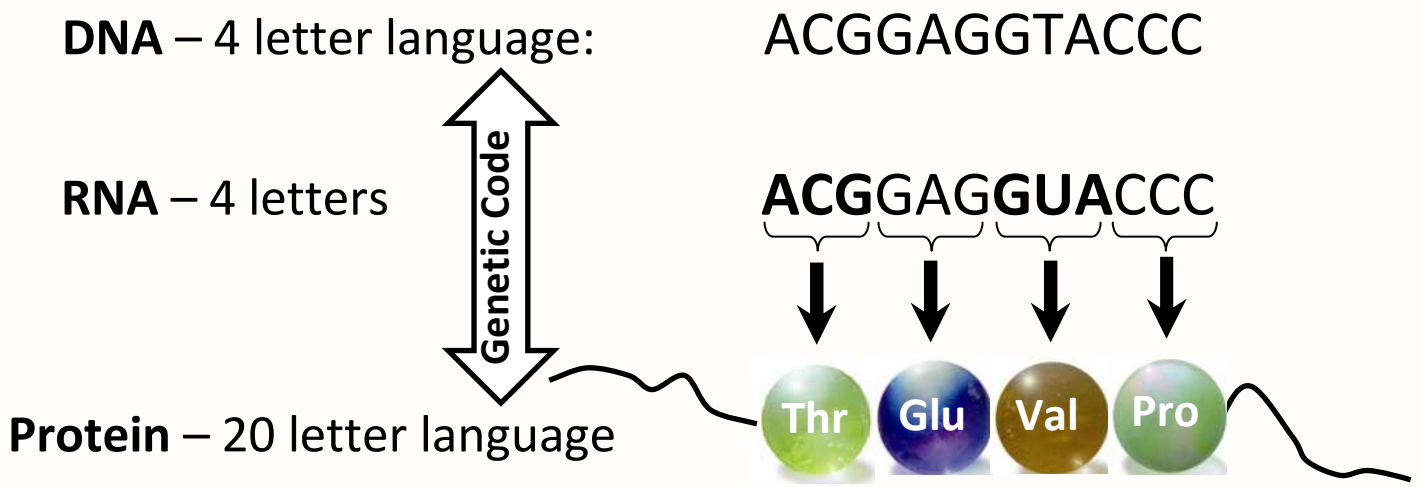
Three Games:

(1) Molecular codes (the genetic code).

(2) Molecular recognition (recombination).

(3) Chromosome organization.

(1) The genetic code is main info channel of life

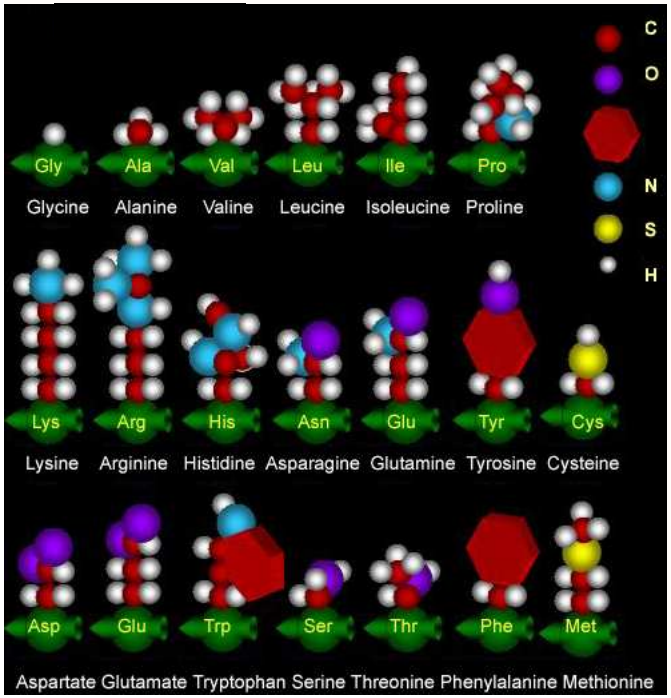


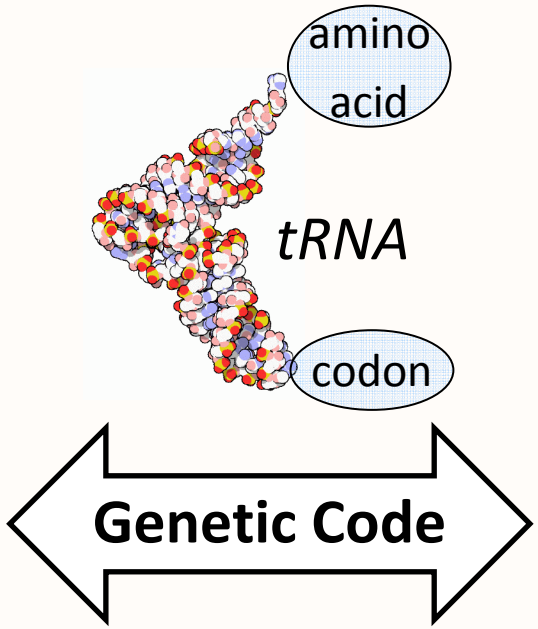
- **Genetic code**: translates 3-letter words in 4-letter DNA language (64 codons) to protein language of 20 amino acids.
- Proteins are amino acid polymers.
- **Diversity** of amino-acids is essential to protein functionality.

The genetic code maps codons to amino-acids

- Molecular code = **map** relating two sets of molecules (spaces, “languages”) via molecular recognition.
- Spaces defined by similarity of molecules (size, polarity etc.)

20 amino-acids



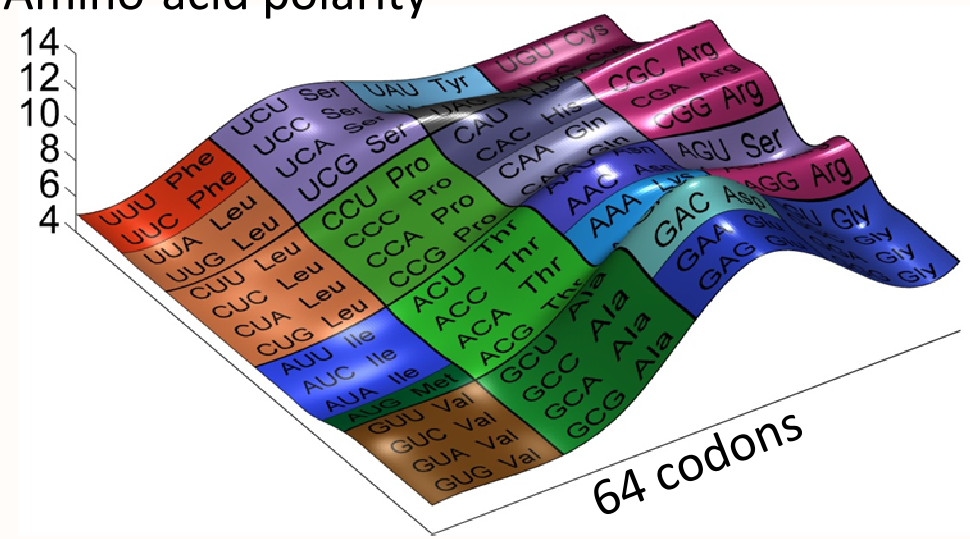


64 codons

The genetic code is a smooth mapping

UUU Phe	UCU Ser	UAU Tyr	UGU Cys
UUC Phe	UCC Ser	UAC Tyr	UGC Cys
UUA Leu	UCA Ser	UAA TER	UGA TER
UUG Leu	UCG Ser	UAG TER	UGG Trp
CUU Leu	CCU Pro	CAU His	CGU Arg
CUC Leu	CCC Pro	CAC His	CGC Arg
CUA Leu	CCA Pro	CAA Gln	CGA Arg
CUG Leu	CCG Pro	CAG Gln	CGG Arg
AUU Ile	ACU Thr	AAU Asn	AGU Ser
AUC Ile	ACC Thr	AAC Asn	AGC Ser
AUA Ile	ACA Thr	AAA Lys	AGA Arg
AUG Met	ACG Thr	AAG Lys	AGG Arg
GUU Val	GCU Ala	GAU Asp	GGU Gly
GUC Val	GCC Ala	GAC Asp	GGC Gly
GUA Val	GCA Ala	GAA Glu	GGA Gly
GUG Val	GCG Ala	GAG Glu	GGG Gly

Amino-acid polarity



- Degenerate (20 out of 64).
- Compactness of amino-acid regions.
- **Smooth** (similar “color” of neighbors).

- Generic properties of molecular codes?

Challenges of molecular codes: rate and distortion

Distortion

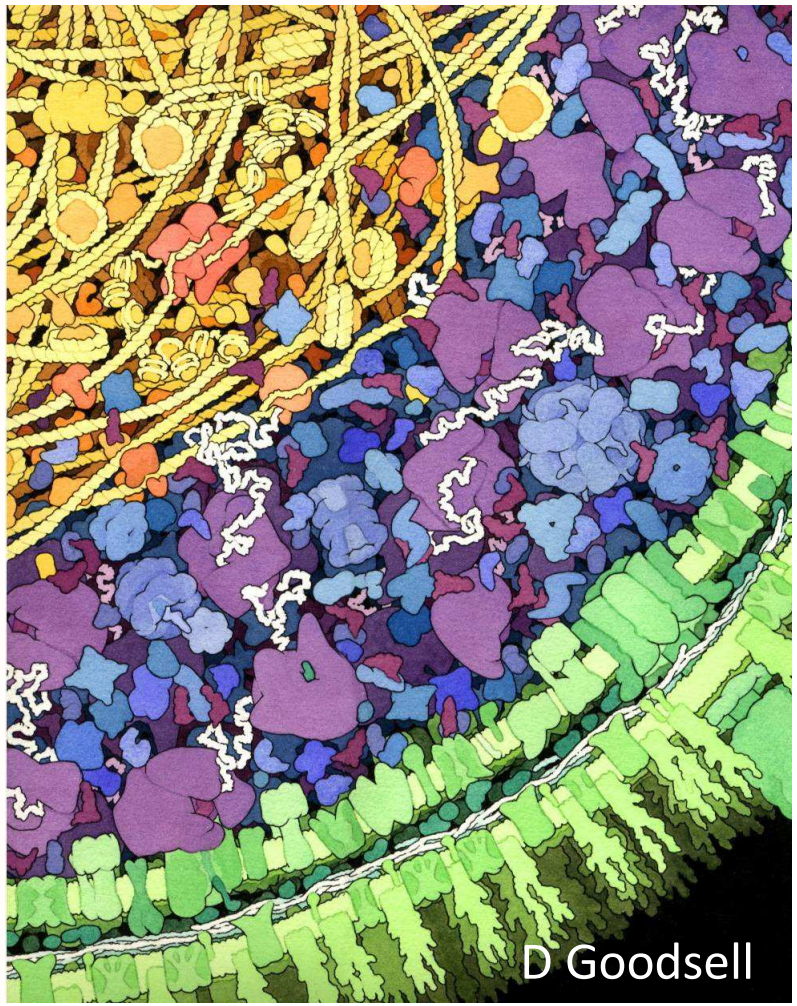
- Recognition in a noisy, crowded milieu.
- Many competing lookalikes.
- Weak recognition interactions ~ $k_B T$.
- Need for diverse meanings.

"Synthesis of reliable organisms from unreliable components" (von Neumann, *Automata Stud.,* 1956)

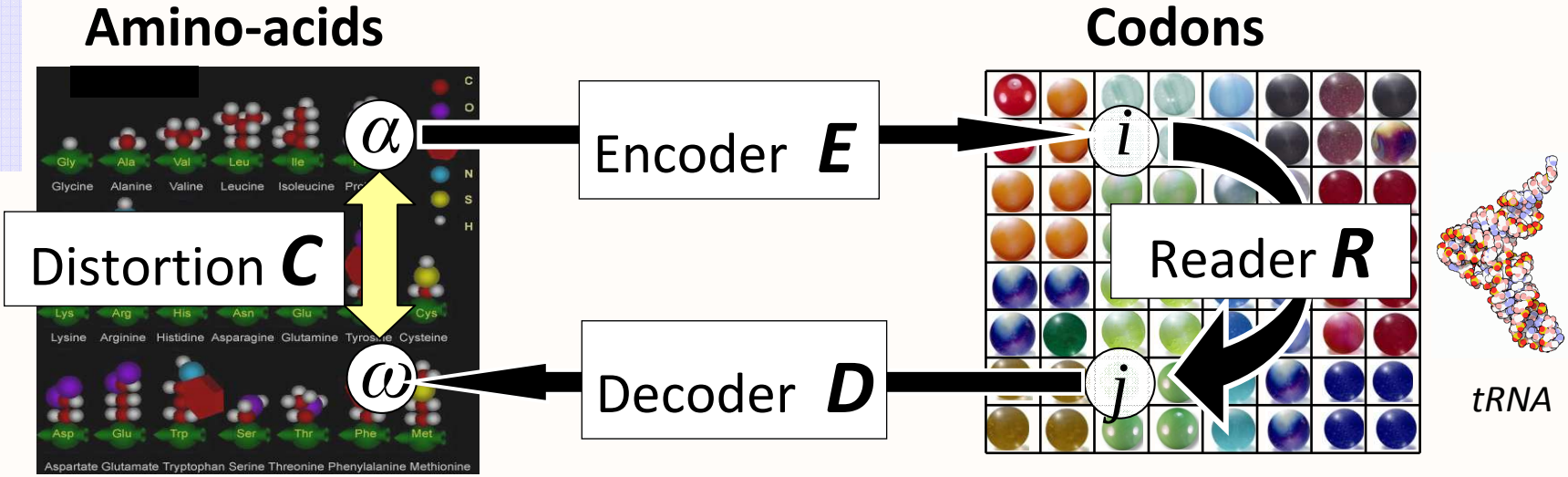
Rate

- How to construct the low-rate molecular codes at minimal cost of resources?

Rate-distortion theory (Shannon 1956)



Distortion determines code's quality

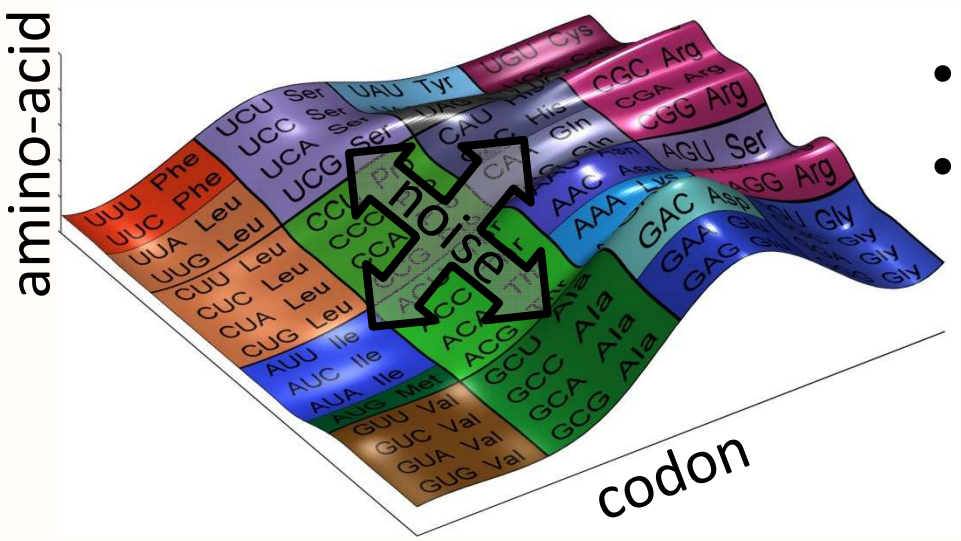


$$Q = \langle C_{\alpha\omega} \rangle = \sum_{paths} P_{path} C_{\alpha\omega} = \mathrm{Tr}\left(E \cdot R \cdot D \cdot C \right)$$

- Distortion of noisy channel ***Q*** = average distortion of AA.
- ***R*** defines topology of codon space.
- ***C*** defines topology of amino-acid space.

(TT, J Theo Bio 2007, PRL 2008, PNAS 2008)

Smooth codes minimize distortion

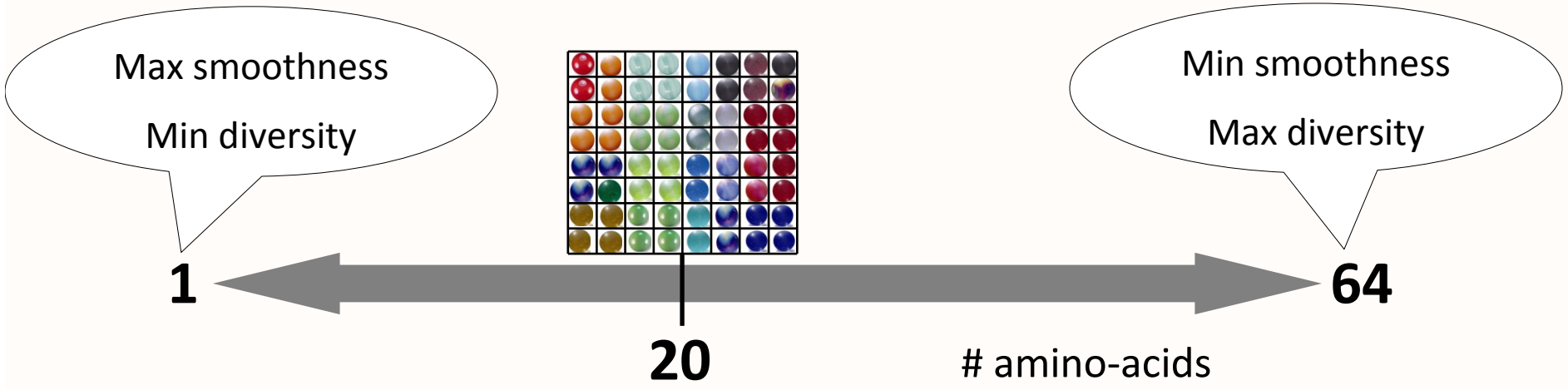


- Errors (noise) confuse close codons.
- Smooth code:

  close codons = close amino-acids.

  → minimal distortion.

- Optimal code must balance contradicting needs for **smoothness** and **diversity**.

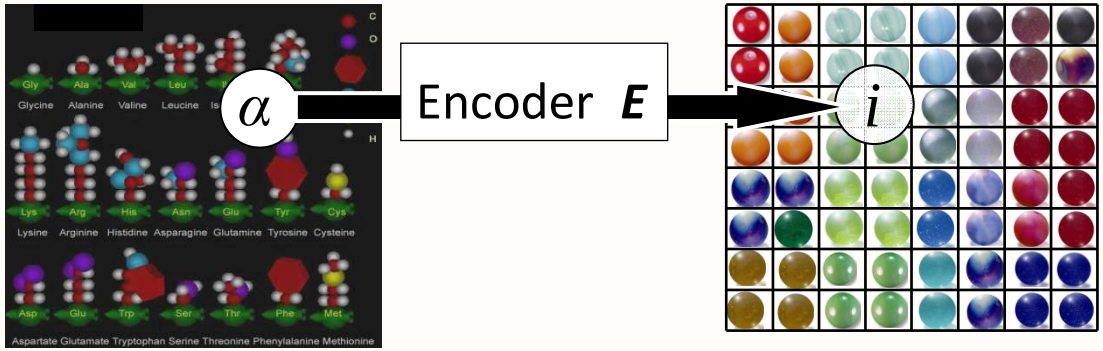


(TT, Bio Phys 2008)

Channel rate is code's cost

- Diverse codes require high **specificity** = high binding energies ε.
- Cost ~ average binding energy $< \varepsilon >$.
- Binding prob. ~ Boltzmann: $\boldsymbol{E} \sim e^{\varepsilon/T}$.

$$I \sim \sum_{\alpha,i} E_{\alpha i} \ln E_{\alpha i} \sim \left\langle \varepsilon_{\alpha i} \right\rangle_E$$



- Cost ***I*** = **Channel Rate** (bits/message)

Code's fitness combines rate and distortion of map

Fitness = Gain x *Distortion + Rate*

$$H = \beta Q + I$$

- ***Gain*** β increases with organism complexity and environment richness.
- Fitness ***H*** is "free energy" with inverse "temperature" β.
- Evolution controls and varies the gain β.

- Population of self-replicators evolving according to code fitness *H*: mutation, selection, random drift.

(TT, PRL, Bio Phys 2008)

Code emerges at coding transition

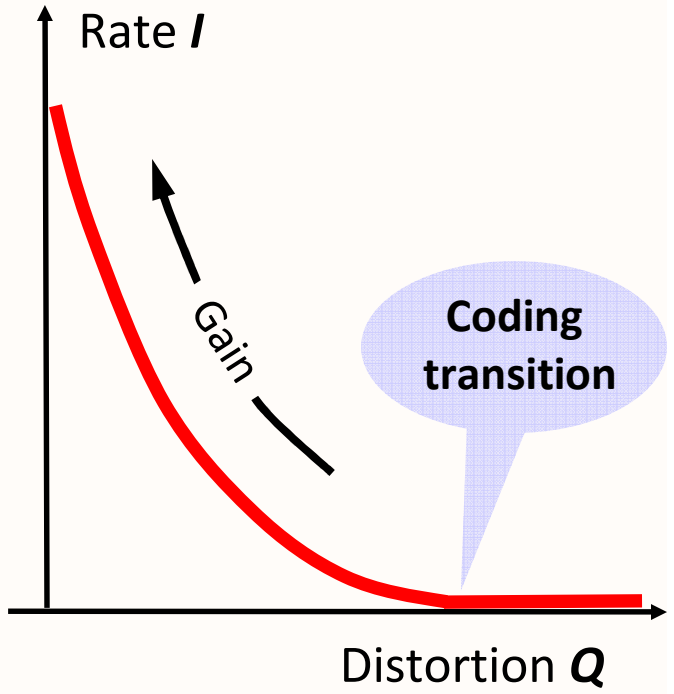
- Low gain β : Cost too high

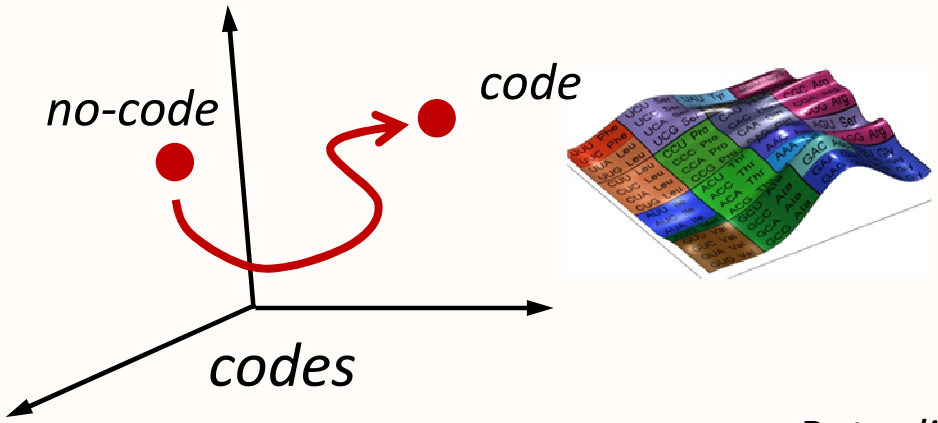
  $\rightarrow$ no specificity $\rightarrow$ **no code**.

- **Code emerges** when β increases:

  channel starts to convey information ($I \neq 0$).

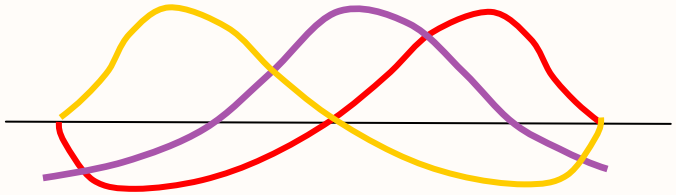
- Continuous phase transition.
- Emergent code is smooth, low mode of ***R***.





Rate-distortion theory (Shannon 1956)

Emergent code is a smooth mode of the error-Laplacian on the symbol graph



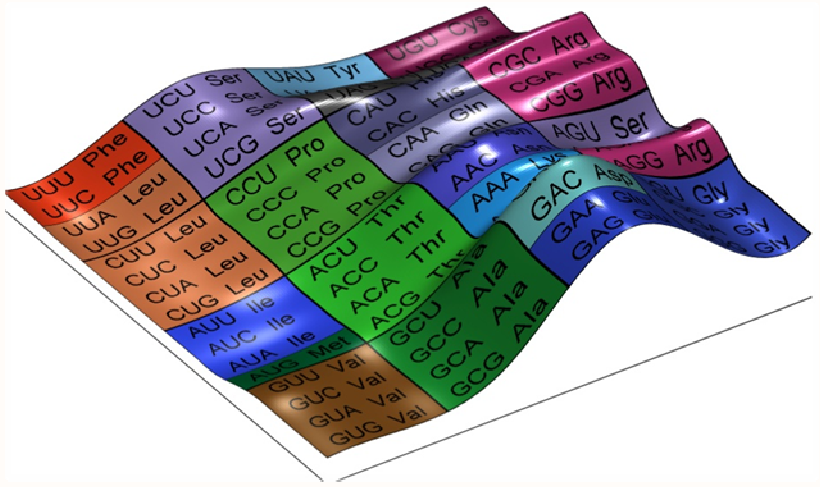
- Every mode corresponds to amino-acid :

  # low modes = # amino-acids.

- Lowest excited modes of graph-Laplacian *R* .
- Single maximum for lowest excited (long wavelength) modes (Courant).

  → single contiguous domain for each amino-acid.

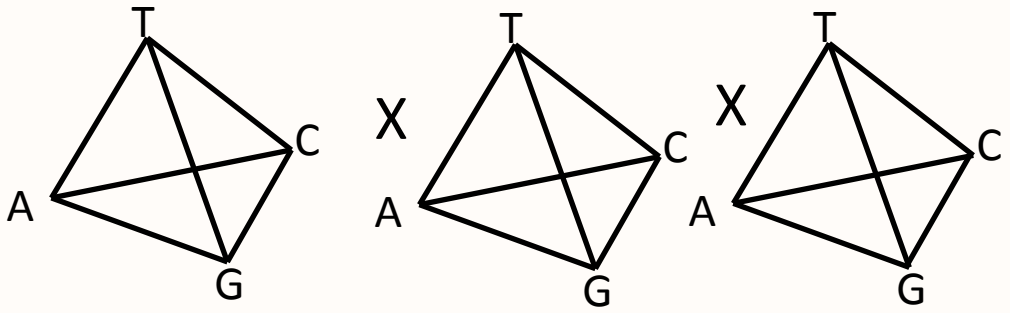
  → **Smoothness**.

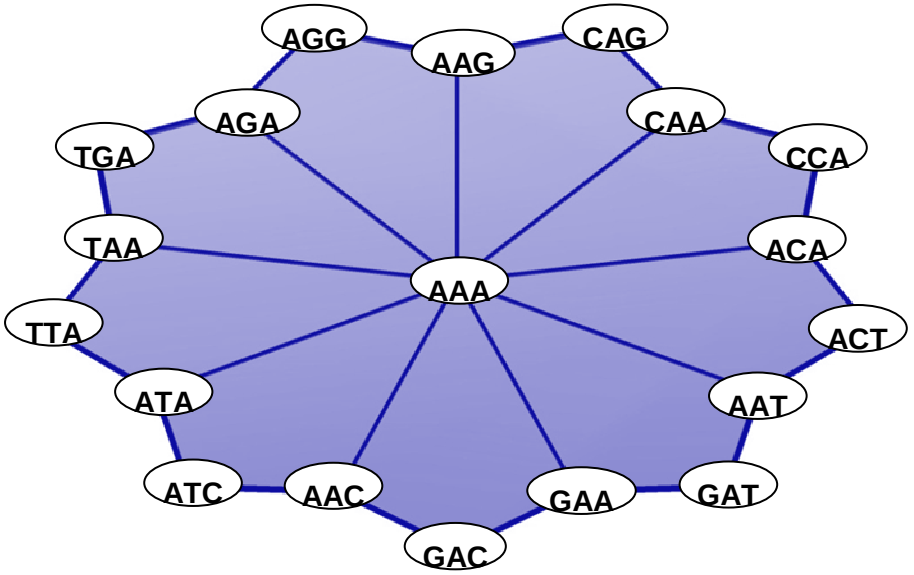


Probable errors define the graph and the topology of the genetic code

- Codon graph = codon vertices + 1-letter difference edges (mutations).

$K_4 \times K_4 \times K_4$





- Non-planar graph (many crossings).
- Genus = # holes of embedding manifold.
- Graph is holey : embedded in $\gamma = 41$

(lower limit is $\gamma = 25$)

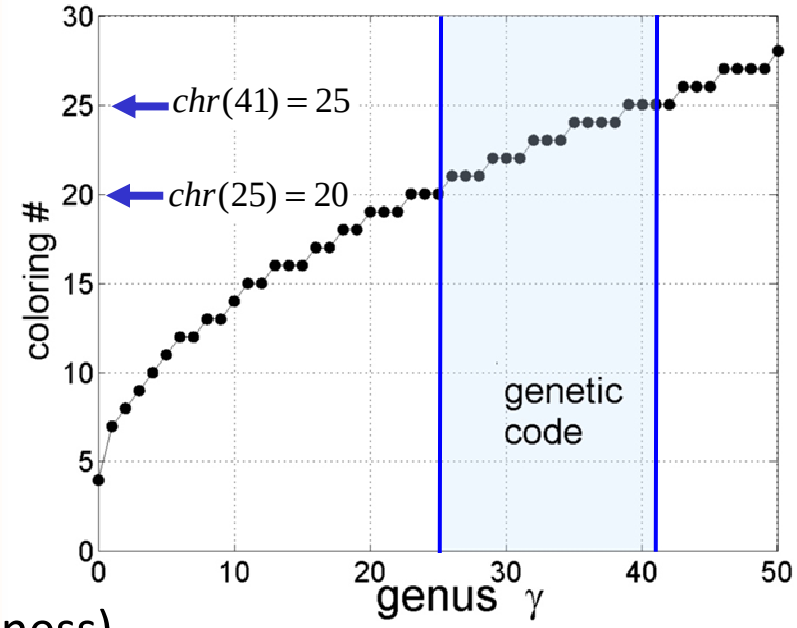
Coloring number limits number of amino-acids

- Q: Minimal # colors suffices to color a map where neighboring countries have different colors?
- A: Coloring number, a topological invariant (function of genus):

$$chr(\gamma) = \left\lfloor \frac{1}{2}\left(7 + \sqrt{1+48\gamma}\right) \right\rfloor.$$

(Ringel & Youngs 1968)

$$\max(\#\text{ amino-acids}) = chr(\gamma)$$



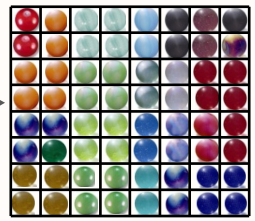
- From Courant 's theorem + "convexity" (tightness).
- Genetic code: γ = 25-41 $\rightarrow$ coloring number = 20-25 amino-acids

(TT *J Lin Alg* 2008)

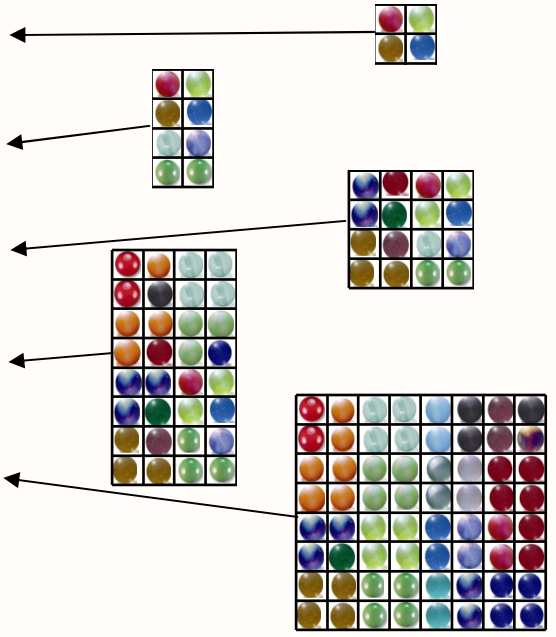
The genetic code coevolves with increasing accuracy of translation

- A path for evolution of codes: from early codes with higher codon degeneracy and fewer amino acids to lower degeneracy codes with more amino acids.

Channel
Mapping
Rate
Distortion
Fitness
Transition
Topology



1st	2nd	3rd	γ	*chr #*
1	4	1	0	4
2	4	1	1	7
4	4	1	5	11
4	4	2	13	16
4	4	3	25	20
4	4	4	41	25



Yoni Savir (WIS)

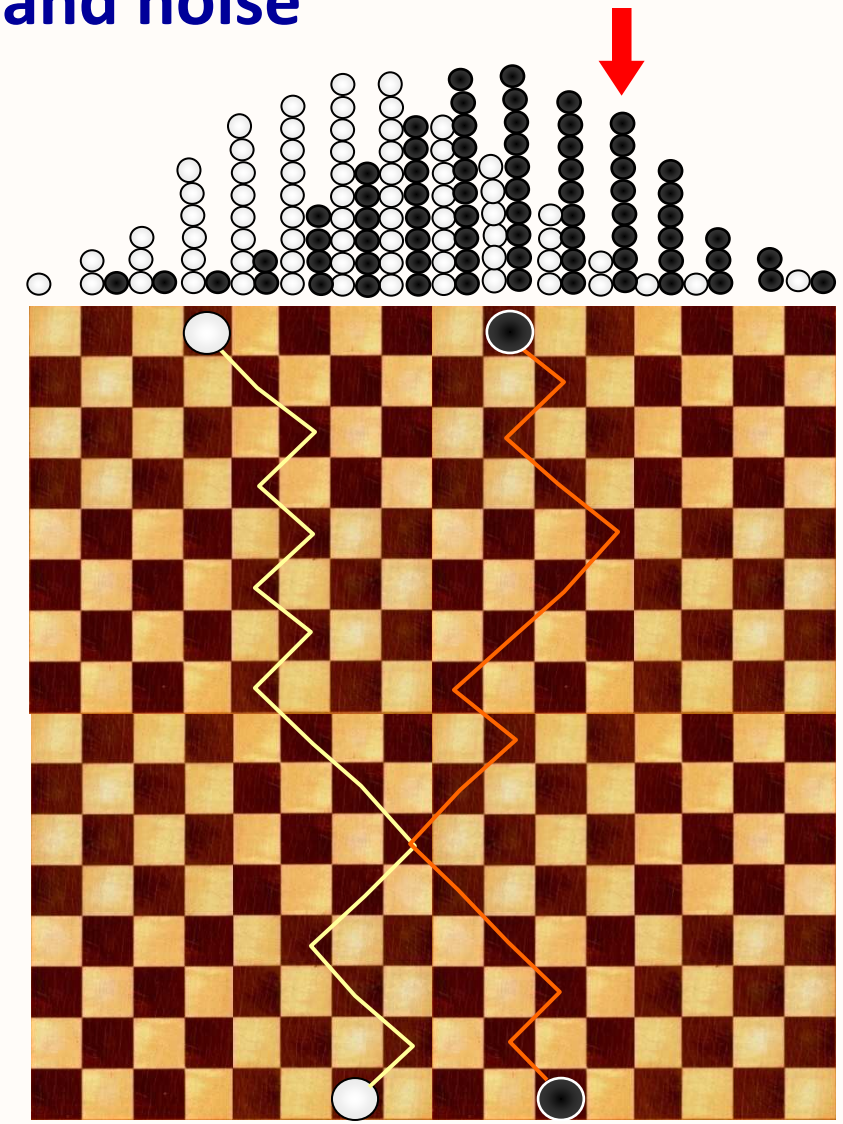
(2) Molecular recognition in presence of competition and noise

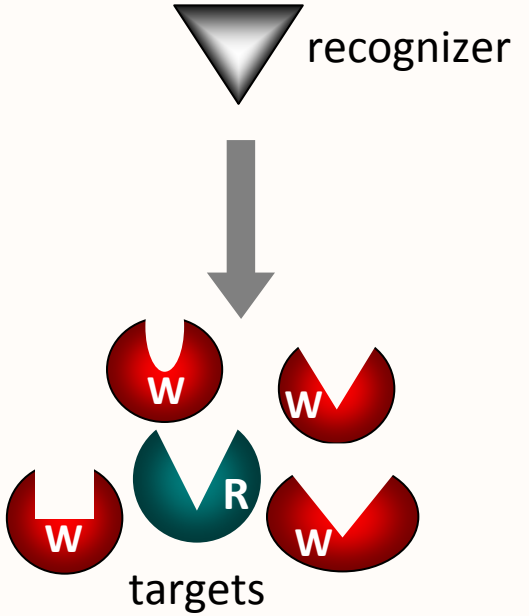
"checkerboard pachinko"

(a) Move B/W checkers up (right or left)).

(b) Collect B/W checkers at the top.

(c) Find max #B - #W.





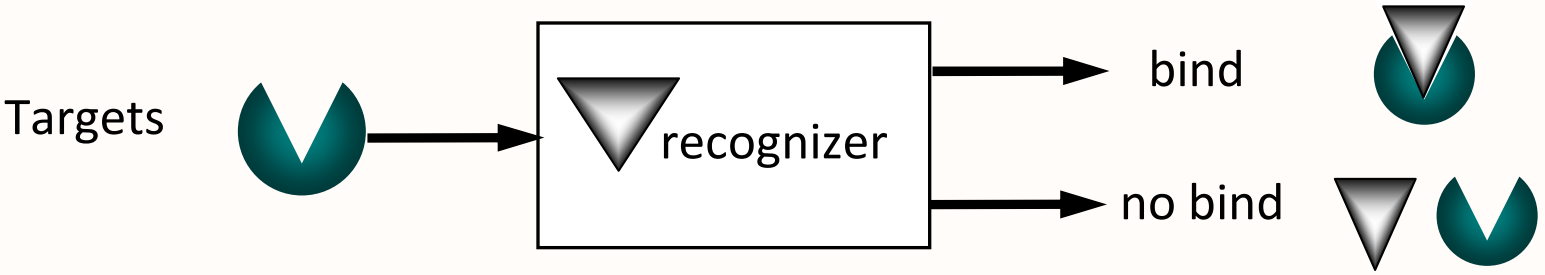
Channel
Mapping
Rate
Distortion
Fitness
Transition
Topology

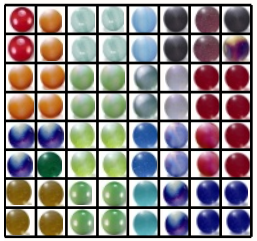
Optimizing noisy molecular information channels

- Info channel: mapping from target space to binding probability:

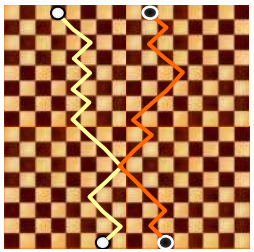
{Targets} ⟶ [0, 1]

- Problem: Optimization with respect to noise.





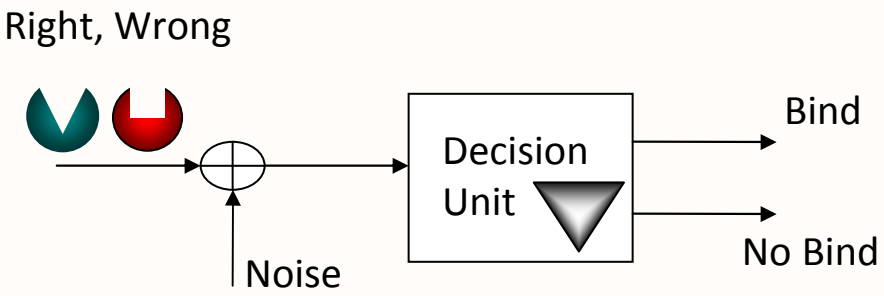
- Large-scale: Assigning outputs to inputs to minimize distortion (marble packing game).



- Small-scale: Improving **accuracy** of each recognition event by tuning structure and conformational change (checkerboard pachinko).

Molecular recognition as a decision problem

- Each (non-) binding event has a benefit (penalty) of (mis-) recognition.
- Performance measure: sum benefit/penalty over binding events.



Decision \ Target	Right	Wrong
Bind	Correct	False Alarm
No Bind	Miss	Correct

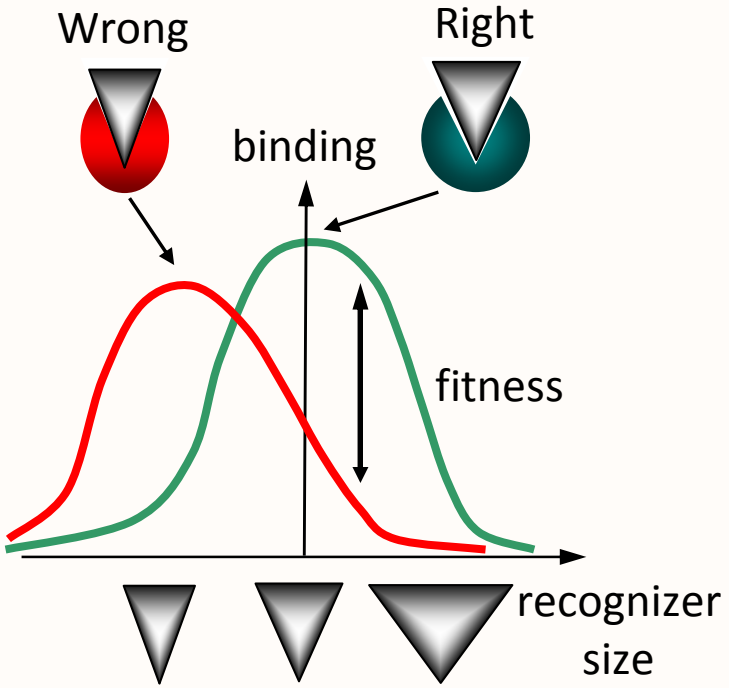
- **Fitness** = Benefit(event)×Prob(event) = correct + miss + false-alarm.
- Fitness depends on structural parameters and can be optimized.

(Savir & TT, Plos 1 2007, IEEE 2008)

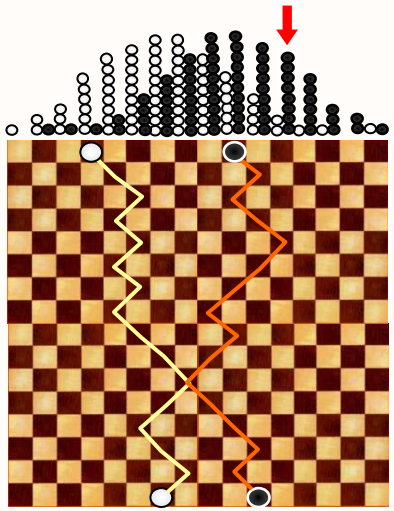
Conformational Proofreading: When off-target is right on

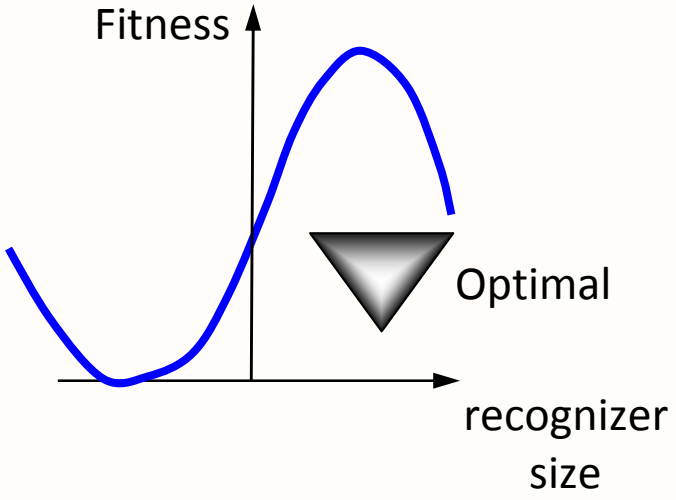
- Structural *mismatch* reduces Right, *but* also reduces Wrong even more.
- Result: Enhancement of Fitness.
- Optimal fitness at non-zero mismatch.

- Optimal recognizer is *off-target*
- Not lock-and-key (*induced fit*)



- Quantitative example: *Homologous Recombination*.

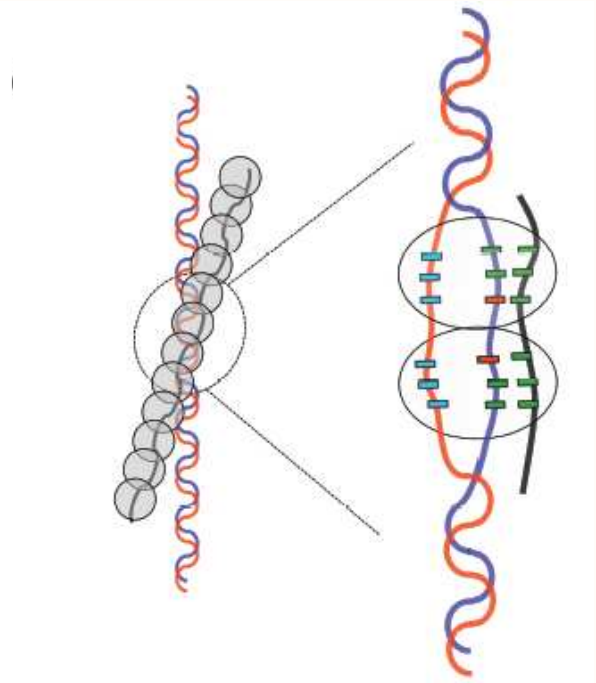




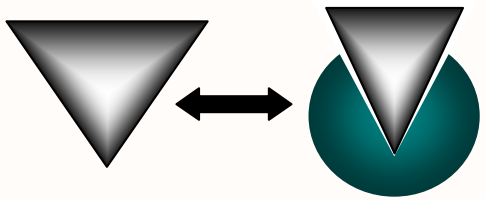
Kinetic proofreading (Hopfield and Ninio 1975/6)

Why DNA is extended in homologous search?

- Recombination = crossover of homologue DNAs.
- Requires **extension** by 50% which costs 3-4 k_BT /bp (k_BT=0.6 kcal/mol).
- Structural reason: exposing bases?



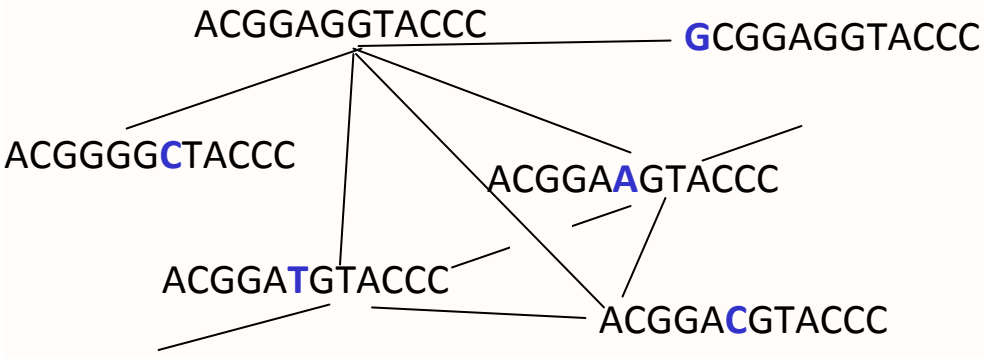
- **Conformational Proofreading**? mechanism to detect Right DNA in large pool of similar targets.

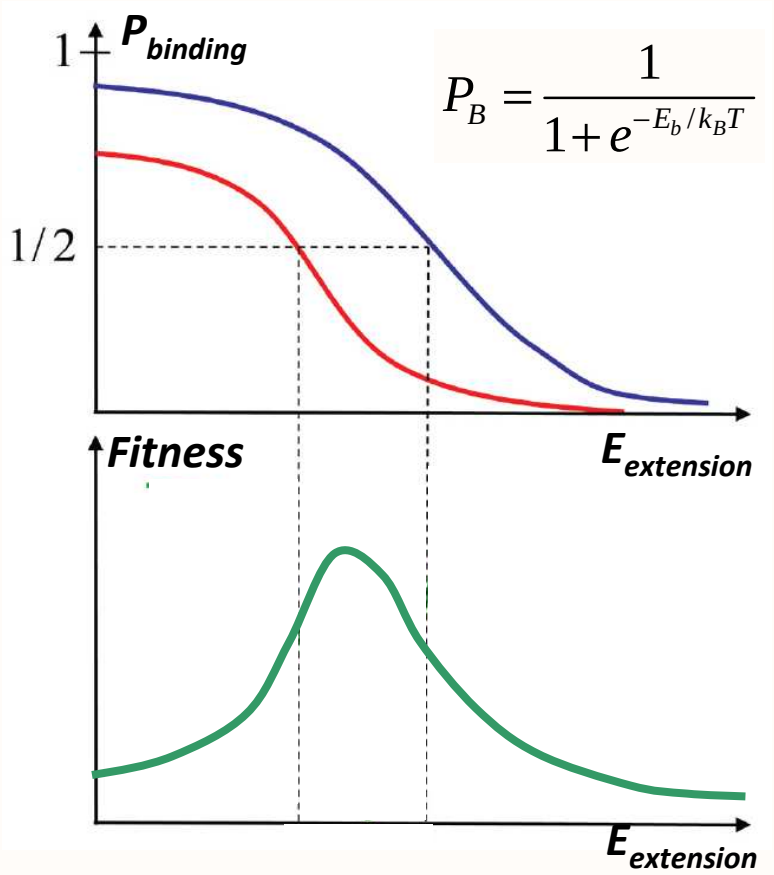


Extension maximizes channel fitness

- Shift in binding energy.
- Extension energy shifts binding to optimal fitness.
- Observed value ~ optimal extension.

- Transition: zero -> non-zero shift.
- Depends on topology of target space.





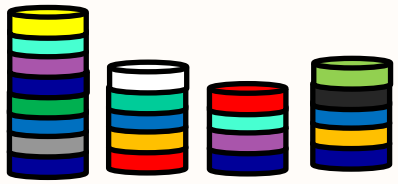
$$P_B = \frac{1}{1+e^{-E_b/k_BT}}$$

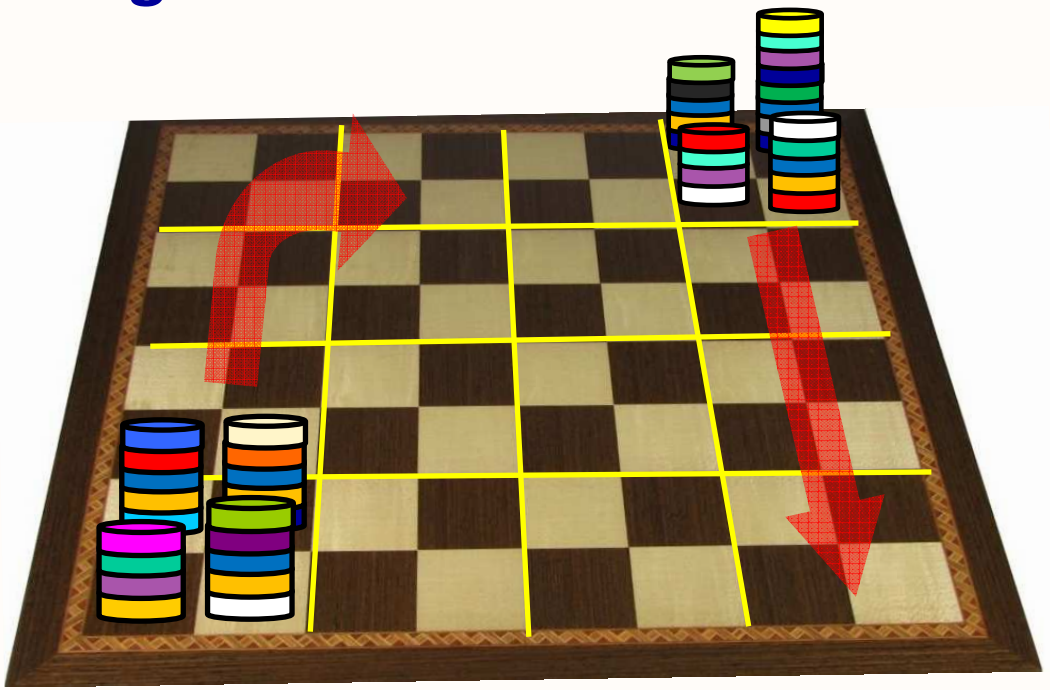
$E_b \sim 1.5\, k_B T / bp$ (Singleton et al., 2006)

$E_{mis} \sim 2.5$ (Camerini-Otero et al., 2006)

E_{ext} Force-extension (Bustamenta al., 2001)

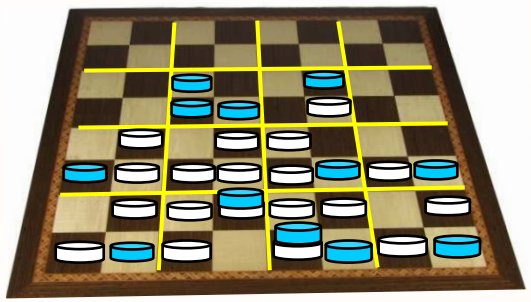
(3) Chromosome Organization

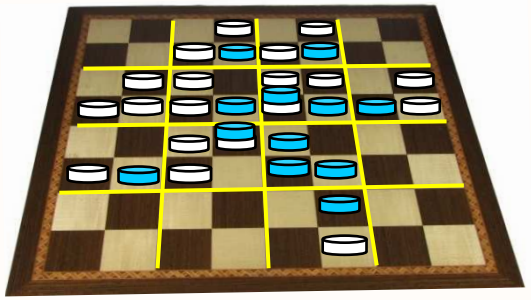




"Casino chip shuffling"

(A) Given stacks of multi-color Chips.
(B) Divide chessboard into "Blocks".
(C) Cover board with stacks.
(D) For each color: Shuffle the Blocks such that
all chips of this color will be close as possible to each other.





A Libchaber (Rockefeller)

GV Shivashankar (NCBS)

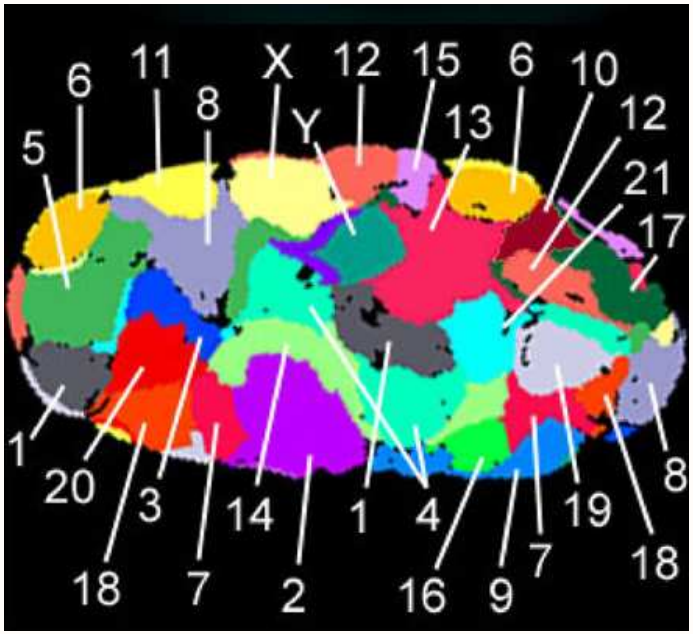
S Maharana

KV Iyer

J.-P. Eckmann

The problem of chromosomes

- Genes are divided into chromosomes.
- What is the optimal (?) number?
- What is the optimal (?) organization?
- Relation to cell type?



Nucleus of G0 human fibroblast (Cremer et al., Plos Bio 2005)

Organism	# chr
M. pilosula (ant)	2
Fruit fly	8
Arabisdopsis	10
C. elegans	12
Rye	14
Corn	20
Chinese hamster	22
Budding yeast	32
Earthworm	36
Cat	38
Syrian Hamster	44
Human	46
Tobacco	48
Silkworm	56
Horse	64
Dog	78
Goldfish	100
Adder's tongue	1400

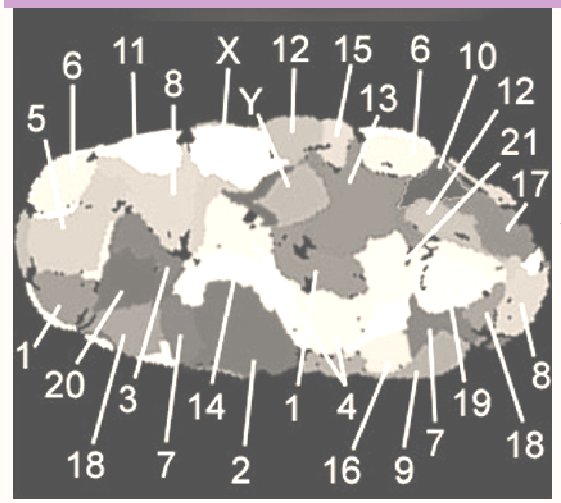
Mapping between 3D organization and expression

Real Space

Cell Types

Function Space

Spatial organization

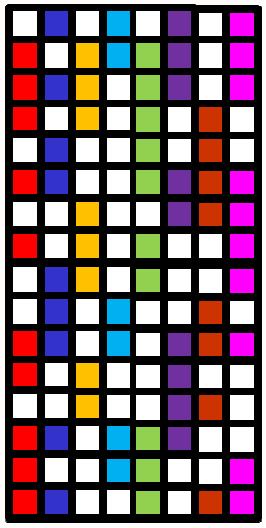


?

Expression

Cell type

genes

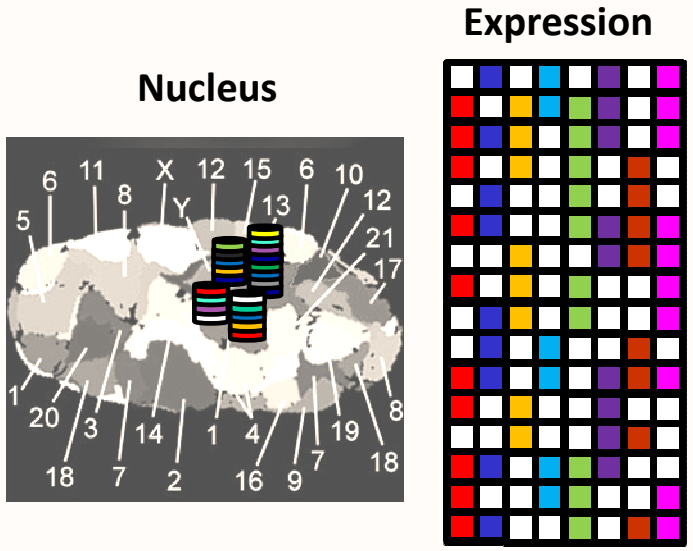


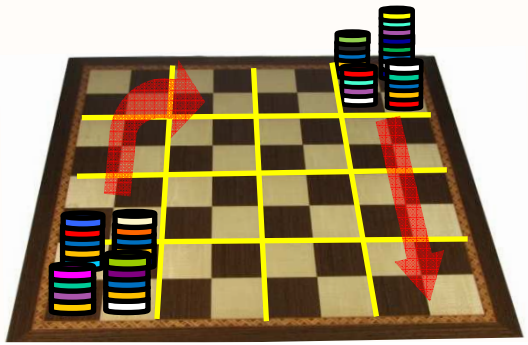
Hypothesis: relation between real space and function space based on optimality

Optimal chromosome organization?

Hypothesis: co-expressed genes or active genes with similar function tend to reside in the same or in close chromosomes.

Optimal organization	Casino chip shuffling
Cell type/function	color
Gene	Chip stack
Chromosome	Block
Nucleus	Chessboard
Chrom. reorganization	Block shuffling
Close active genes	Close Same color chips





Inter-chromosomal interactions

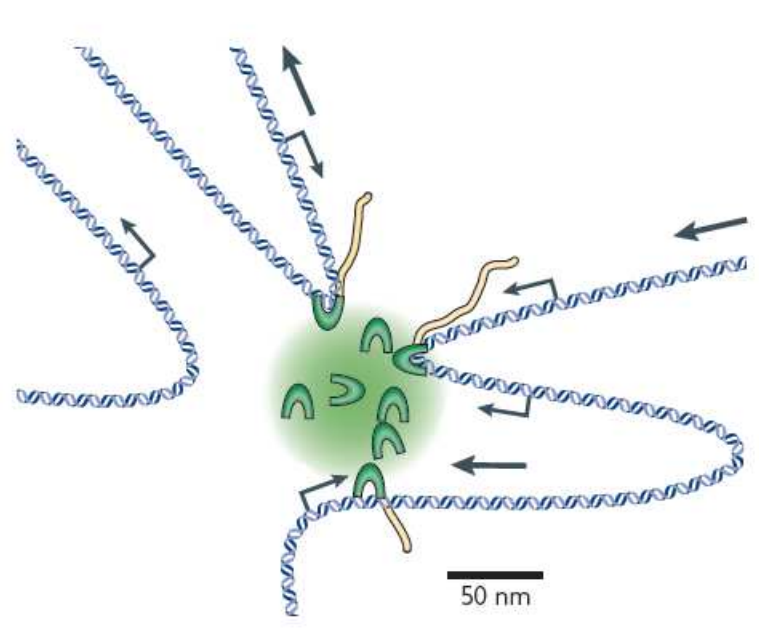
Expression space

- Genetic networks, co-regulation, co-expression.

Physical space

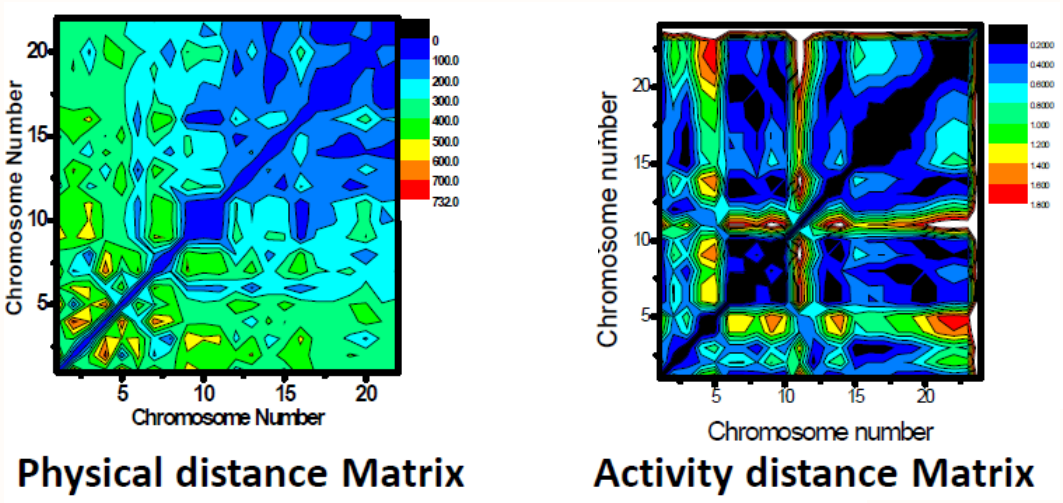
- Physical proximity boosts efficiency of transcription factories?
- Small nuclear RNA (snRNA)?

  small nuclear ribonucleoproteins (snRNP)?
- ….
- ***Smoothness***



Transcription factories: Genes from same or from different chromosomes may associate with polymerases in the same factory. (Sutherland & Bickmore, *Nat Rev Gen* 2009)

Correlation between organization and expression

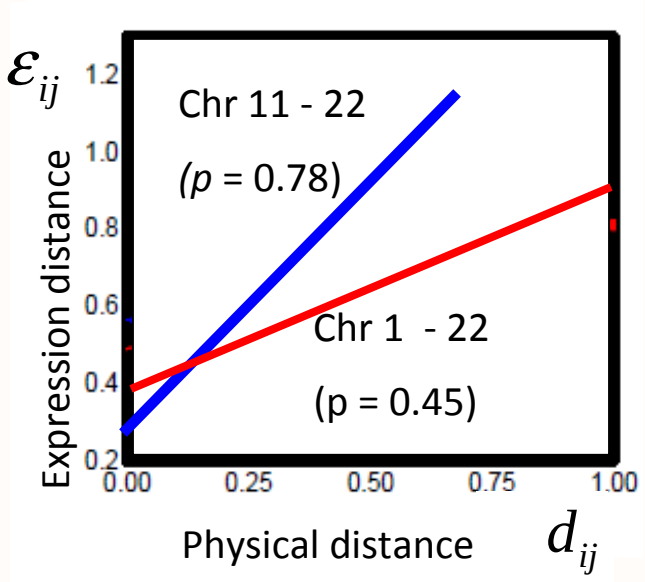


- Expression distance

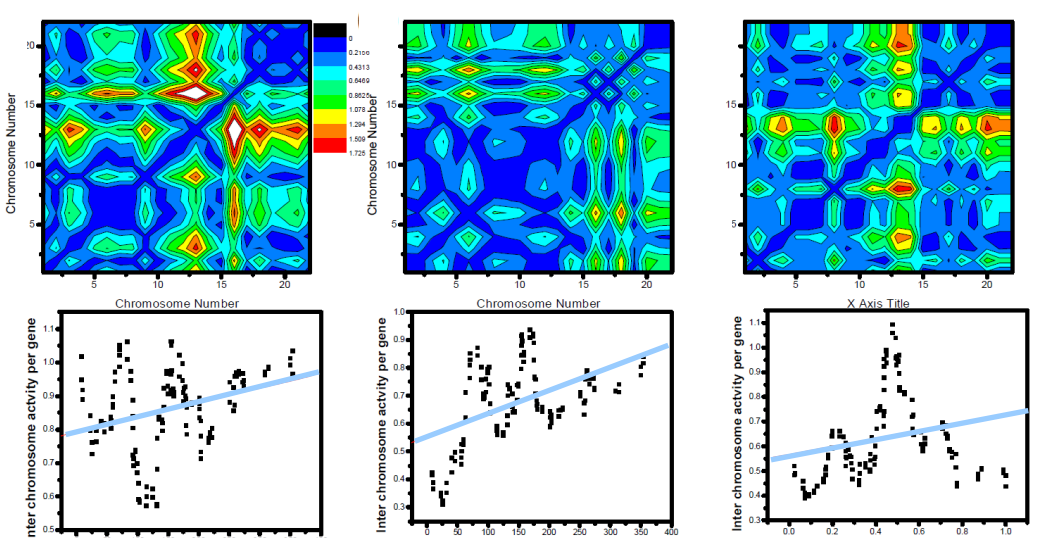
$$\varepsilon_{ij} = \left| \ln \frac{\langle \text{activity/gene} \rangle_i}{\langle \text{activity/gene} \rangle_j} \right|$$

- Physical distance

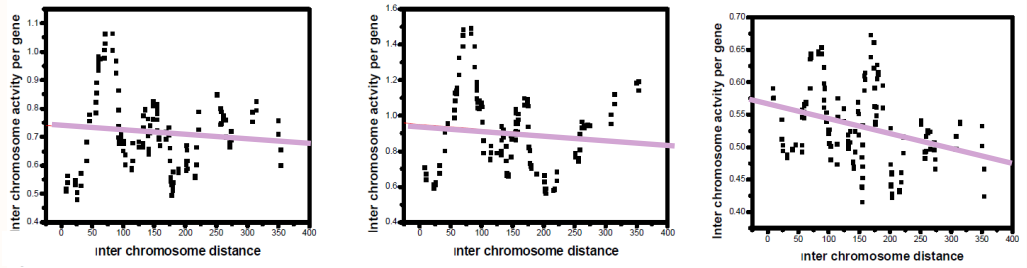
$$d_{ij}$$

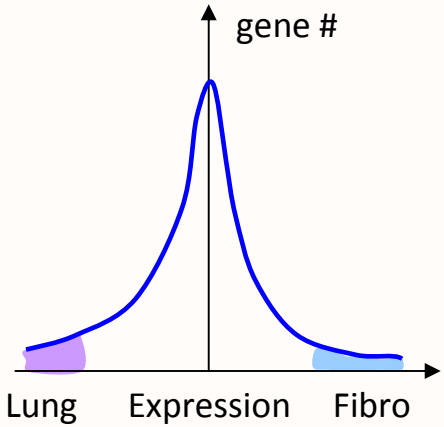


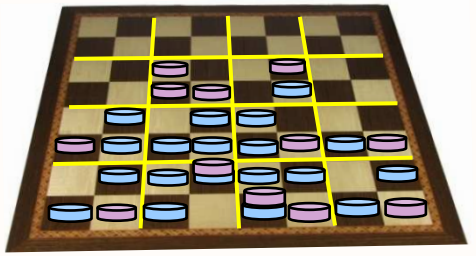
Is chromosome organization cell specific?



Lung vs. Fibroblast Oocyte vs. Fibro HUVEC vs. Fibro







Human Fibroblast Cells : GSM157869 Goetze et al., Mol Cell Bio 2007.
Lung Cells : GSM101102 airway epithelial (Cornell Med.,2009)
Oocytes : GSM288812 pooled, unfertilized mature (Montpellier, 2009)
HUVEC : GSM215557 (Johns Hopkins, 2009)

HUVEC - human umbilical cord vein endothelial cell.
Oocyte – female germ cells.
Fibroblast – connective tissue.
Lung – epithelial cells.

Measure for optimality of chromosome organization

- "Smoothness" measure (similar to genetic code's distortion *Q*)

  How close are gene that perform the same function in given cell type.

$$H^C = \sum_{chr\ ij} A_{ij}^C \sum_{nets\ n} B_n^C \left(P_{ni} - P_{nj}\right)^2$$

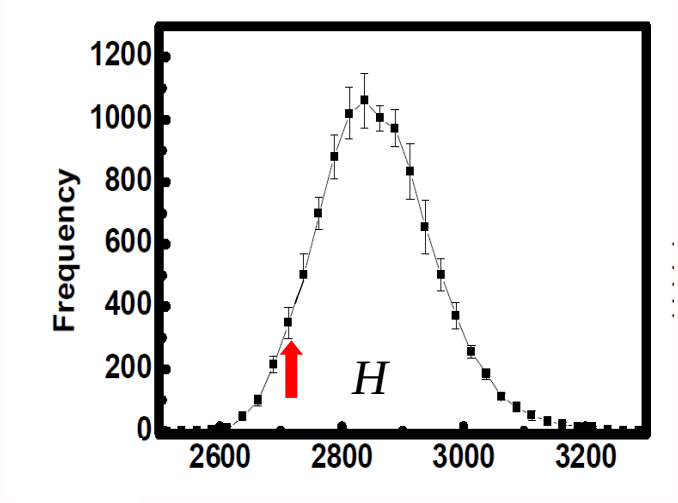
in cell type C:

A_{ij}^C = "adjacency" matrix $\sim 1/d_{ij}$

B_n^C = is netwrok n active (0,1)

P_{ni} = activity of nework n in chrom. i

- To test optimality: calculate H for shuffled chromosomes.



$$H_{observed} \simeq \langle H_{random} \rangle - 2 \cdot \sigma$$

p-value < 0.001

Questions and directions

- Better optimality measures (transcription factories).
- Other cell types (preliminary evidence from T cells).
- Optimality transition as a function of chromosome #.
- Relations to the topology of the chromosome graph (A_{ij}).

- Other molecular information channels: molecular recognition, transcription networks (U. Alon S. Itzkovitz).

Itzkovitz Alon & TT (BMC Genomics 2006); Shinar TT Alon (PNAS 2006)

Channel
Mapping
Rate
Distortion
Fitness
Transition
Topology

THANKS

Albert Libchaber (RU)

Yoni Savir (WIS)

Uri Alon (WIS)

Shalev Itzkovitz

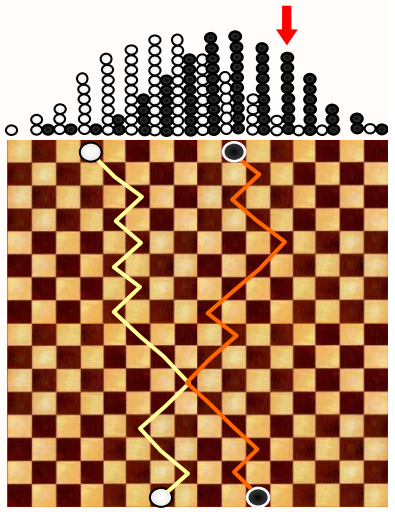
Guy Shinar

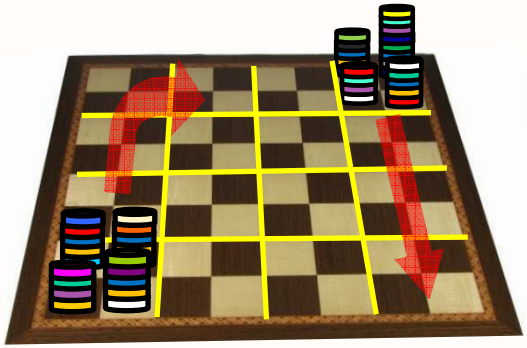
G.V. Shivashankar (NCBS)

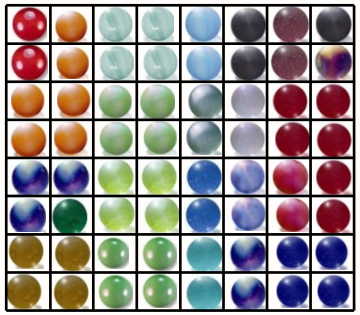
Shovamayee Maharana

K.Venkatesan Iyer

J.-P. Eckmann (Geneva)







Papers: www.weizmann.ac.il\complex\tlusty